Annotated Documents and Expanded CIDOC-CRM Ontology in the Automatic Construction of a Virtual Museum

Cristiana Araújo, Ricardo G. Martini, Pedro Rangel Henriques and José João Almeida

Abstract The Museum of the Person (Museu da Pessoa, MP) is a virtual museum with the purpose of exhibit life stories of common people. Its assets are composed of several interviews involving people whose stories we want to perpetuate. So the museum holds an heterogeneous collection of XML (eXtensible Markup Language) documents that constitute the working repository. The main idea is to extract automatically the information included in the repository in order to build the virtual museum's exhibition rooms. The goal of this paper is to describe an architectural approach to build a system that will create the virtual rooms from the XML repository to enable visitors to lookup individual life stories and also inter-cross information among them. We adopted the standard for museum ontologies CIDOC-CRM (CIDOC Conceptual Reference Model) refined with FOAF (Friend of a Friend) and DBpedia ontologies to represent OntoMP. That ontology is intended to allow a conceptual navigation over the available information. The approach here discussed is based on a TripleStore and uses SPARQL (SPARQL Protocol and RDF Query Language) to extract the information. Aiming at the extraction of meaningful information, we built a text filter that converts the interviews into a RDF triples file that reflects the assets described by the ontology.

C. Araújo (✉) · R.G. Martini · P.R. Henriques · J.J. Almeida
Department of Informatics, Algoritmi Research Centre, University of Minho,
4710-057 Gualtar, Braga, Portugal
e-mail: decristianaaraujo@hotmail.com

R.G. Martini
e-mail: rgm@algoritmi.uminho.pt

P.R. Henriques
e-mail: prh@di.uminho.pt

J.J. Almeida
e-mail: jj@di.uminho.pt

Á. Rocha and L.P. Reis (eds.), *Developments and Advances in Intelligent Systems and Applications*, Studies in Computational Intelligence 718,
DOI 10.1007/978-3-319-58965-7_7

1 Introduction

The society is more and more concerned with the preservation and the dissemination of Cultural Heritage, as works of art, ancient objects, and documents, among others.

Nowadays this can be achieved in a better way resorting to the information and communication technologies because they allow that the physical objects, on one hand, become accessible to anyone, and on the other hand, are not deteriorated rectos [1–3].

In this context of technological expansion, increasing the capability of extraction, storage and visualization of everyday life events, the museums have taken advantage to expand its field of action, as well as their own concept. They expand their geographical borders by providing information in their pages on the Internet and exhibiting their collections. On the other hand, completely virtual environments (called Virtual Museums, VM) appeared, without any references to physical spaces [2].

A Virtual Museum, such as a traditional museum, also acquires, conserves, and exhibits the heritage of humanity (in that case, intangible objects, or immaterial things[1]) creating a delightful environment for pleasure or enjoyment, as well as an appropriate place for teaching, and research.

This article is concerned with the creation of a specific Virtual Museum, the Museum of the Person (MP). The assets of the MP contains several interviews that narrate the life stories of ordinary citizens. These citizens, to report their life stories, remember events and other particular situations they have participated in. MP resources are constituted by a collection of documents in XML (eXtensible Markup Language) format.

In the article we discuss the interest and the way of building a virtual museum (that we see was a virtual learning space) to tell to the world those life stories and to extract knowledge about an epoch and a society connecting and relating them.

More precisely we aim at rebuilding npMP, the Portuguese branch of the Museum of the Person network (this network includes branches in Brazil, Portugal, USA, Canada, etc.) that connect individuals and groups through sharing their life stories (http://www.museumoftheperson.org/about/).

In this paper, and after a brief introduction to MP (Sect. 1.1), we discuss the ontology built to describe the museums knowledge repository (Sect. 2), then we present different technical approaches to implement the desired virtual museum (Sect. 3) and, finally, we introduce and describe the first module of our system that extracts required information from XML repository and its storage in the triple store that instantiate the ontology (Sect. 4).

Besides OntoMP, an ontology for the museum of the person that is new and a first contribution of this work, also the extension of the standard CIDOC-CRM for museums with FOAF and DBpedia concepts and properties is another contribution presented. The discussion on DBpedia inclusion is new material not yet presented in previous conference version of this article.

[1] According to: http://www.unesco.org/culture/ich/index.php?lg=en\Źpg=00022#art2.

An important contribution of our work presented in the paper is the detailed definition of a generic architecture for the implementation of a system that creates the museum exhibition rooms from the documents repository. Moreover we designed and propose two possible implementations of that generic architecture, one more appropriate for situations where the repository is stored in a relational database, and the other to be used when the repository is archived in a triple-store. Our aim is to compare both approaches to understand the development effort involved in each one and to learn their benefits and drawbacks.

At the best of our knowledge there are not similar projects that use ontologies and tools to generate automatically virtual learning spaces from their specifications, neither in the scope of MP nor in the context of other virtual museums. So we will not include a section on related work. For the sake of space (necessary to introduce all the novelties of this paper) we decided not to include a state of the art section; the reader is referred to the authors pre-thesis [4, 5], where we review the form main topics: Ontologies and CIDOC-CRM; Cultural Heritage; Learning Spaces; and Virtual Museums.

1.1 Museum of the Person, an Overview

Museum of the Person aims at gathering testimonials from every human being, famous or anonymous, to perpetuate his history [1, 3].

Life stories are evidences in support of facts or statements attested by common people carrying a social and historical character, which must be preserved and processed to become an immeasurable human heritage (intangible or immaterial things). The interviewed are used as informers, reporting the events and emotions they experienced [1].

To report their life stories during a predefined structured interview, the narrators remember events and other particular situations they have participated in. These memories will act as a basic element for social research [1].

The Museum of the Person's collection consists of sets of XML documents, specified by a DTD (Document Type Definition created specially for that purpose and called MP-DTD) related with each participant. Typically each interview is split into three parts [6]:

- **BI:** a brief biography and personal data, such as name, date and place of birth, and job;
- **interview:** two versions of the interview are built and saved—the *interview* file refers to the raw interview and contains all the questions asked and the narrator's answers; the *edited* file is a plain text, structured by themes that define small portions of a person's life story. In this format, a life story may give rise to thematic stories (e.g., dating, childhood, craft, among others). Both *interview* and *edited* files contain metadata tagging;
- **photographs and their caption**. This caption includes a description of the image, people depicted, place and the date.

Aside the interviews, there is also a *thesaurus* that includes key concepts mentioned in the stories.

Details about the elements that constitute each DTD will be mentioned in the next section that will discuss the development of MPs ontology (OntoMP). For more details on Museum of the Person please see [7].

2 The CIDOC-CRM Ontology for MP, OntoMP

2.1 OntoMP: Original Design

After an exhaustive analysis of all the documents (XML instances, respective DTD's, and the thesaurus) that belong to Museum of the Person, we could identify the concepts and relations involved in the life stories. This first step enabled us to design OntoMP, an ontology for the Museum of the Person. In this way, the museum visitor can have a conceptual navigation over the collection.

The main concepts extracted from the analysis phase are: people *(pessoa)*, ancestry *(ascendência)*, offspring*(descendência)*, job *(profissão)*, house episode *(episódio casa)*, education episode *(episódio educação)*, dating episode *(episódio namoro)*, general episode *(episódio geral)*, childhood episode *(episódio infância)*, leisure episode *(episódio lazer)*, religious episode *(episódio religioso)*, accident *(evento acidente)*, migration *(migração)*, life's philosophy *(filosofia de vida)*, festivity *(festividade)*, catastrophic event *(evento catastrófico)*, political event *(evento político)*, marriage *(casamento)*, birth *(nascimento)*, dream *(sonho)*, uses *(costumes)*, religion *(religião)* [7, 8].

In a similar way we also identified the following relations: performs *(exerce)*, depicted *(éRetratada)*, visits *(visita)*, lives *(vive)*, receives *(recebe)*, tells *(narra)*, has *(tem)*, has-type *(tipo)*, enrolls *(participa)*, occurs *(ocorre)*, refers to *(dizRespeito)* [7, 8].

Then we realized that some more elements should be added to the ontology. The concepts added were: marital status *(estadoCívil)*, spouse *(cônjuge)*, widowhood *(viuvez)*, sex *(sexo)*, literacy *(habilitações literárias)*, political party *(partido político)*, first communion *(primeira comunhão)*, death *(morte)*, baptism *(batismo)*, child's birth *(nascimento do filho)*, photos *(fotos)*, description *(descrição)* and file *(ficheiro)* [7, 8].

The ontology so far obtained is depicted in Fig. 1.

Figure 1 shows the main concepts in a life story (ellipsis) related with Person and also shows his main data properties (rectangles). Figure 1 enhances Event concept (a relevant component of OntoMP) and its different sorts (subclasses).

To validate the ontology designed, we created some instances using actual life stories picked-up from the MP collection, as can be consulted in the projects site at the http://npmp.epl.di.uminho.pt. Notice that all those interviews were conducted in the past and we got written permissions to publish them.

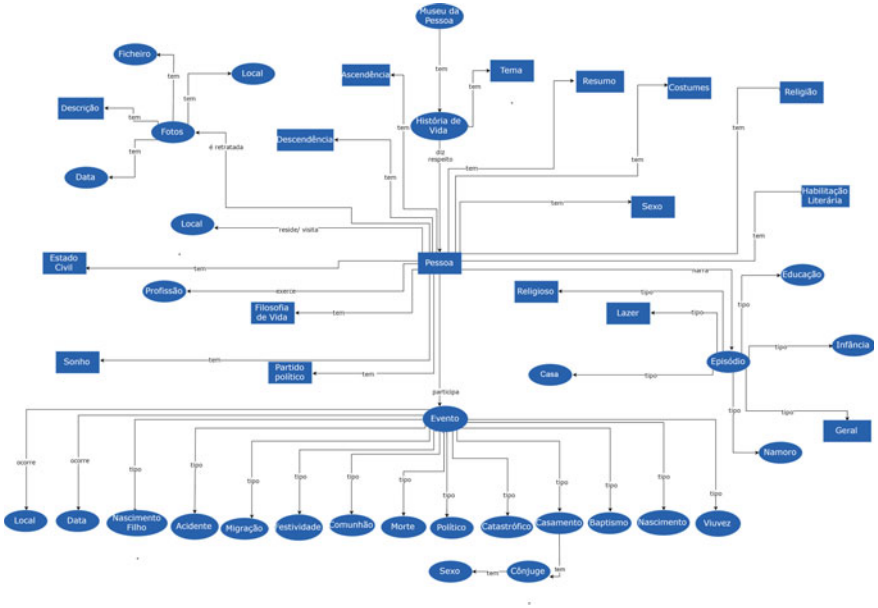


Fig. 1 An ontology *OntoMP: Original Design* for MP

2.2 *OntoMP: CIDOC-CRM/FOAF/DBpedia Representation*

After the validation and tuning of OntoMP, the next stage was to describe it in a standard ontology format used for museums, CIDOC-CRM (CIDOC Conceptual Reference Model). For that purpose we have followed the approach adopted in the context of another project to build Portuguese Emigration virtual museum [9].

CIDOC-CRM is a formal ontology planned to aid in the integration, mediation, and interchange of heterogeneous Cultural Heritage information [10]. It specifies the semantics of museums documentation.

CIDOC-CRM is an Event-based ontology, and therefore it should contain *Time-Spans* and *Places* related with each event. The core of CIDOC-CRM is based on seven concepts: *Temporal Entities*, *Events*, *Actors*, *Time-Spans*, *Conceptual Objects*, *Physical Things*, and *Places*. Notice that, *Actors* and *Conceptual Objects* or *Physical Things* should also be related with *Event* [10].

The transformation of *OntoMP: Original Design* in CIDOC-CRM was a straightforward process; the original concepts were expressed as events and associated concepts, and the original relations were mapped into the correspondent in CIDOC-CRM.

However, we found that some properties related with person could not be expressed in CIDOC-CRM in a simple and natural manner. So we decided to explore the combination with FOAF (Friend of a Friend) and DBpedia, since both contain a vocabulary specific to describe individuals, their activities and their relations with other people and objects [11].

FOAF ontology describes two areas of digital identity information: biographical and social network information [12].

DBpedia ontology is a shallow, cross-domain ontology, which has been manually created based on the most commonly used infoboxes within Wikipedia. DBpedia knowledge base covers various fields, such as geographic information, people, businesses, online communities, movies, music, books and scientific publications, among others [13].

After this investigation, we refined CIDOC-CRM adding some pertinent FOAF and DBpedia concepts and properties. Regarding FOAF, we imported *gender* property, person names (*name*, *givenName*, *familyName* and *nick*) and person-image relations (*depicts* and *depiction*). From DBpedia we picked up properties like *religion*, *profession*, *education*, *party* and *spouse*.

After the refinement of CIDOC-CRM ontology with FOAF and DBpedia elements, we got a simpler notation (descriptions became less verbose); moreover the original was enriched conceptually, this is more details about person's stories can be included in the knowledge base. The final OntoMP represented in this new notation was once again instantiated with concrete data extracted from the real life stories. It was possible to validate it once more.

In Fig. 2 we show an instance of the ontology created with data extracted from Maria Cacheira interview. Below we describe the CIDOC-CRM, FOAF and DBpedia fragment reproduced.

A person *(E21 Person)*, *gender* Female, *name* Maria Alice Rodrigues Cacheira (decomposed in *givenName* Maria Alice and *familyName* Rodrigues Cacheira), *participated in (E5 Event)* that is her birth *(E67 Birth)*. This event occurred at a *(E52 Time Span)*—that *is identified by (P78)* 1946-10-08, an *(E50 Date)*—and at a *(E53 Place)*—that *is identified by (P87)* Afurada an *(E44 Place Appellation)*.

This person *(E21 Person)* is *depicted* in the photo *(E38 Image)*. This photo *is identified by (P1)* 090-F-01.jpg *(E41 Apellation)*, *has note (P3)* Maria Alice Rodrigues Cacheira, *refers to (P67)* Maria Alice Rodrigues Cacheira *(E55 Type Description)*, and was taken in a *(E52 Time Span)*—that *is identified by (P78)* 2001-12-07, an *(E50 Date)*—and at a *(E53 Place)*—that *is identified by (P87)* Junta de Freguesia da Afurada, an *(E44 Place Appellation)*.

A person *(E21 Person)* has *education* "Sabe ler e escrever (4^{a} classe)", professes the *religion* "Católica" and has *profession* "Peixeira e Empregada de limpeza".

In this fragment of Maria Cacheira's life story other concepts can be identified. All these concepts, that characterize a *(E21 Person)*, are represented in CIDOC-CRM version, as *(E55 Type)*. For example, *(E21 Person) has type (P2)* "Viúva" *(E55 Type Marital Status)*.

The person's properties imported from FOAF (above identified) are emphasized in Fig. 2 using dotted line. Similarly, DBpedia properties used are enhanced as dashed line.

This CIDOC-CRM ontology enriched with FOAF and DBpedia elements can describe appropriately the knowledge repository of the Museum of the Person.

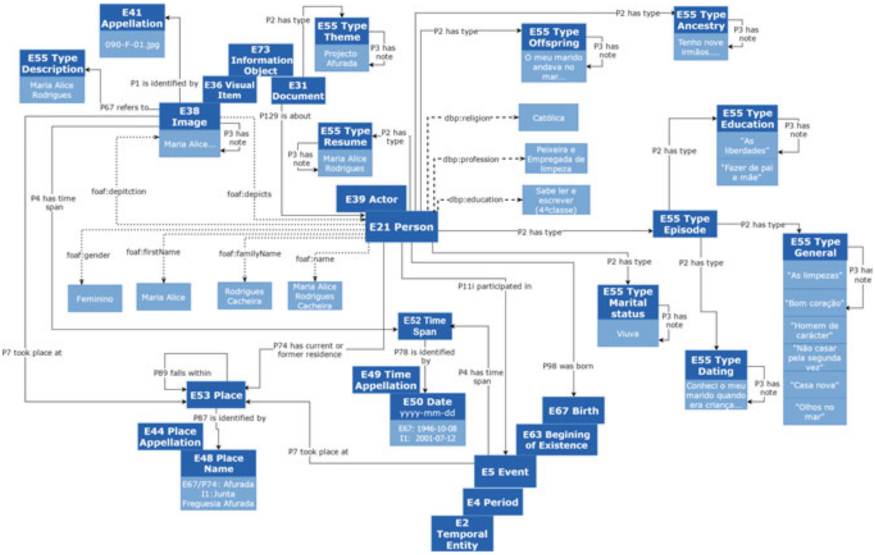


Fig. 2 An instance of CIDOC-CRM/FOAF/DBpedia representation of OntoMP for Maria Cacheira life story (fragment)

3 Proposed Architectures

This section presents a general approach to create a system that builds automatically the Museum of the Person from its repository.

This proposal is defined at an abstract level so that the main architectural blocks and their interactions can be clearly understood; the data flow and the main transformations will be emphasized without technological commitments. We have devised and sketched two possible technical alternatives to implement general architecture. However after describing the general approach, only alternative 1 will be detailed because is the one we chose to refine that architecture.

The general approach, illustrated in Fig. 3, to build the MP comprises: the repository; the Ingestion Function [M1] responsible for getting and processing the input data; a Data Storage (DS) that is the data digital archive; an Ontology to map and link the concepts with the objects stored in (DS); the Generator (M2) to extract data from (DS) and manage the information that will be displayed in Virtual Learning Spaces (VLS) (the final objective of this project) [8].

As said above this general approach has two possible refinements, which are dependent on the (DS). In approach 1 (Figs. 4 and 5) the [DS] is a *TripleStore*, while in approach 2 the (DS) is a *Relational Database*. According to the kind of storage chosen, the ingestion function and the learning spaces generator will require different designs [8].

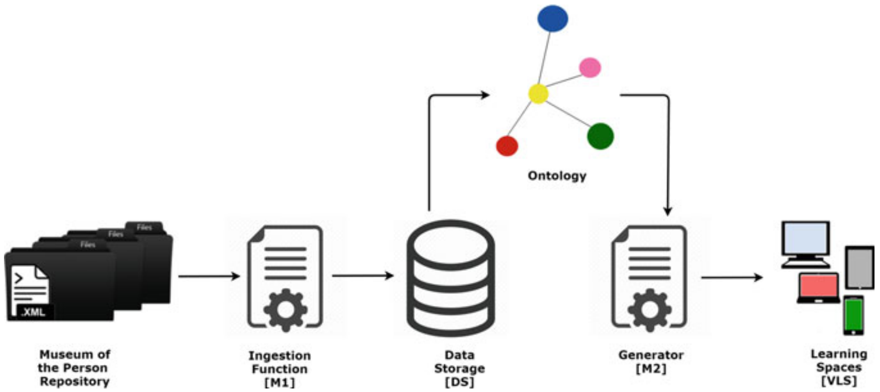


Fig. 3 General Approach to build the MP

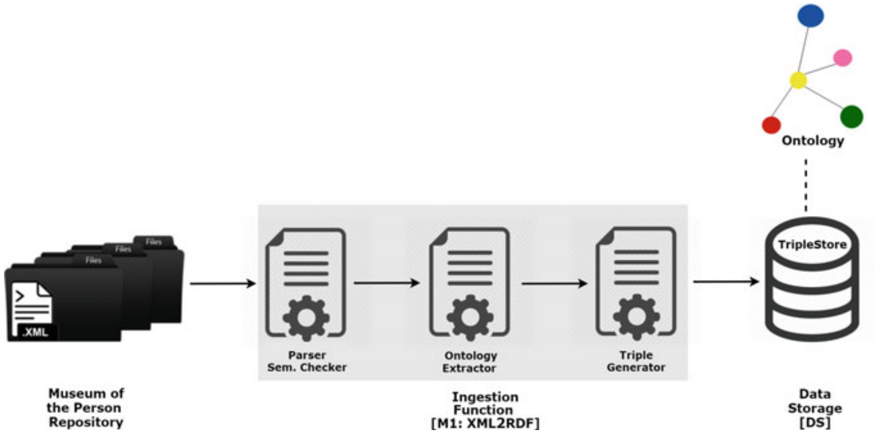


Fig. 4 Module [M1] in Approach 1

The project here imported, we are following the approach 1, so we will describe it in more detail in Sect. 3.1.

In Approach 2 the input XML documents must be converted into SQL to populate the respective database. So, Ingestion Function (M1) is composed of the following components: *Parser and Semantic Checker* that reads the repository documents and extracts the relevant data (annotated in XML), checking their semantic consistency; and *SQL Generator* that generates automatically the SQL statements that insert the retrieved data into the database tables. After the two phases of Ingestion Function [M1] the documents data populate the Relational Database schema, due to the SQL statements generated. As this schema is not directly related to the ontology, in this second approach an explicit mapping is necessary. Making this mapping available, it is possible to resort to CaVa (Criação de Ambientes Virtuais de Aprendizagem) system [4] to build automatically the Virtual Learning Spaces (VLS). Notice that

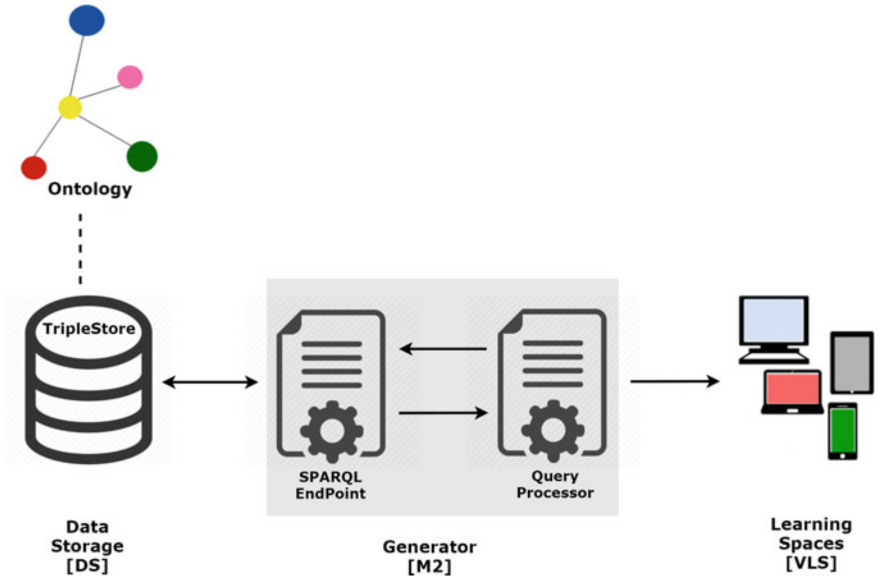


Fig. 5 Module (M2) in Approach 1

only the generator module of CaVa, CaVa^{Gen} will be used in this context. In this case, Generator [M2] is composed of: *DB2Onto Mapping* that associates concepts and relations belonging to the ontology with their respective instances stored in database (it allows to access database tables and fields to get the instances of the ontology concepts); and *CaVa*^{Gen}, that generates automatically the Virtual Learning Spaces from their formal specification based on the ontology.

In this second approach all the work concerned with the query generation according to the exhibition requirements and the answer processing to fulfill the rooms templates is left to CaVa^{Gen}. The only thing that is needed is the specification of the desired learning spaces in CaVa^{DSL}. For more details about this approach please read [8]. The next section will detail approach 1.

3.1 Approach 1

As said above, Approach 1 is based on the decision of using a TripleStore as data storage (DS). According to this decision, Ingestion Module and the Generator Module must be adapted; the first will transform the input XML (eXtensible Markup Language) documents into RDF (Resource Description Framework) triples, and the second will retrieve information from the RDF triples to create the museum web pages.

Figure 4 details the first module that is composed of three blocks [8]:

- *Parser and Semantic Checker* that reads the repository documents and extracts the relevant data (annotated in XML), checking their semantic consistency;
- *Ontology Extractor* that identifies in the extracted data the concepts and relations that belong to the ontology creating in this way an instance of the abstract ontology (in another words, this component populates the ontology);
- *Triple Generator* that converts automatically the ontology triples (created in the preceding block) into triples in RDF notation appropriated to be stored in the (DS) chosen.

At an early stage, to realize the kind of information that contains the documents and how we would represent, we decided to conduct the analysis and extraction of information from documents manually. This means that we accomplished the three phases of Ingestion Module manually.

Among the many existing notations for describing ontologies we chose RDF because we use CIDOC-CRM, FOAF and DBpedia that are described in its original form in RDF. An excerpt of the RDF triples built by hand is shown in Listing 1.

Listing 1 Fragment of the RDF Triples for Maria Cacheira life story

```
<!-- Description Interviewed 1 -->
<rdf:Description rdf:about="&ecrm;Interviewed_1">
        <rdf:type
   rdf:resource="&ecrm;E21_Person"/>
        <rdf:type rdf:resource=
   "http://dbpedia.org/ontology/Person"/>
        <rdf:type rdf:resource="&foaf;Person"/>

        <foaf:firstName
   rdf:datatype="&xsd;string">Maria
   Alice</foaf:firstName>
        <foaf:name
   rdf:datatype="&xsd;string">Maria Alice
   Rodrigues Cacheira</foaf:name>
        <foaf:familyName
   rdf:datatype="&xsd;string"> Rodrigues
   Cacheira</foaf:familyName>
        <P98i_was_born rdf:resource="&ecrm;B1"/>

        <foaf:gender rdf:datatype="&xsd;string">
   Feminino</foaf:gender>
        <foaf:depiction rdf:resource=
   "&ecrm;I1_Interviewed_1"/>

        <dbp:profession
   rdf:datatype="&xsd;string"> Peixeira e
   empregada de limpeza</dbp:profession>
        <dbp:religion rdf:datatype="&xsd;string">
   Catolica</dbp:religion>
        <dbp:education
   rdf:datatype="&xsd;string"> Sabe ler e escrever
```

```
        (quarta classe)</dbp:education>
      </rdf:Description>
18
19  <!-- Event Birth Interviewed 1  (B1) -->
20  <rdf:Description rdf:about="&ecrm;B1">
21          <rdf:type rdf:resource="&ecrm;E67_Birth"/>
22          <P98_brought_into_life rdf:resource=
      "&ecrm;Interviewed_1"/>
23          <P4_has_time-span
      rdf:resource="&ecrm;TS1"/>
24          <P7_took_place_at
      rdf:resource="&ecrm;PL1"/>
25  </rdf:Description>
26
27  <!-- Description Photo Interviewed 1 (I1) -->
28  <rdf:Description
      rdf:about="&ecrm;I1_Interviewed_1">
29          <rdf:type rdf:resource="&ecrm;E38_Image"/>
30          <rdf:type rdf:resource="&foaf;Image"/>
31          <foaf:depicts
      rdf:resource="&ecrm;Interviewed_1"/>
32          <P67_refers_to rdf:resource=
      "&ecrm;I1_Description_Interviewed_1"/>
33          <P3_has_note rdf:datatype="&xsd;string">
      Maria Alice Rodrigues Cacheira</P3_has_note>
34          <P1_is_identified_by rdf:resource=
      "&ecrm;090-F-01.jpg"/>
35          <P4_has_time-span
      rdf:resource="&ecrm;TS7"/>
36          <P7_took_place_at
      rdf:resource="&ecrm;PL8"/>
37  </rdf:Description>
```

The triple fragment shown in Listing 1 contains information about life story of Maria Cacheira. The biographic information about Maria Cacheira, as name (first, last name and full name), birth, sex, photo, profession, religion, and education is displayed in first section (line 1–17). The birth event of Maria Cacheira, date and place of it, is described in the second section (line 19–25). Finally, the last section (line 27–37) contains specific information about the photo of the interviewed, such as description, legend, file, date and place.

The next step was to use the W3C online tool RDF Validator[2] to validate the handwritten triples to ensure that the very long textual description produced contains no errors. RDF Validator checks the consistency of the triple RDF and displays them in a table with three columns ‘subject, predicate and object’. After loading our RDF file we got, as feedback, the information *“VALIDATION RESULTS: Your RDF document validated successfully”* that is just what we want to get from that tool.

The next step, after the successful validation, was to store the triples in a data set, a RDF database, called Apache Jena TDB.

[2] https://www.w3.org/RDF/Validator/.

TDB is a component of Jena (free and open source Java framework for building Semantic Web and Linked Data applications) for RDF storage and query, and can be used as a high performance RDF store on a single machine. A TDB store can be accessed and managed with the provided command line scripts and via the Jena API. Apache Jena Fuseki component provides a SPARQL server to be used with TDB [14].

By performing these three phases of the Ingestion Function (M1), we understand how to make the extraction and analysis of semantic concepts and how to convert the triple ontology in RDF triples. You also realize that it is a very time consuming work to be done manually for all documents in the repository study, then we decide to create a tool to do these three phases automatically. This tool will be described in detail in Sect. 4.

As the mapping between the domain ontology (previously defined) and the data extracted from the repository is automatically built by construction in the second block, above, there is no need to create explicitly this mapping. It means that the Generator [M2] can access directly the storage to obtain the conceptual information necessary to create the exhibition rooms [8].

To display in the Virtual Learning Spaces (VLS) the information stored in (DS)–TripleStore, the (VLS) Generator needs to send queries and process the returned data.

Figure 5 shows the second module [M2] (the Generator) that is composed of two blocks [8]:

- *SPARQL Endpoint* that receives and interprets the SPARQL queries, accesses the TripleStore and returns the answers. For this, it is necessary to resorted to a SPARQL Endpoint. The SPARQL endpoint used was Apache Jena Fuseki (version 2.0).

  Apache Jena Fuseki is a SPARQL server, that can run as an operating system service, as a Java web application (WAR file), and as a standalone server. Fuseki is tightly integrated with TDB to provide a robust, transactional persistent storage layer, and incorporates Jena text query and Jena spatial query [15]. To check if we could extract information from the created ontology, we built some queries. An example of a query that has been built is the one listed below to find the name of the Interviewed of a given sex and residence.

Listing 2 Query SPARQL: Interviewed by sex and residence

```
PREFIX : <http://erlangen-crm.org/150929/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dbp: <http://dbpedia.org/ontology/>
PREFIX rdf:
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT DISTINCT ?name

WHERE {

```

```
11    ?pessoa a :E21_Person;
12      :P129_is_subject_of ?doc;
13      foaf:name ?name;
14      foaf:gender"sexInterviewed"^^xsd:string ;
15      :P74_has_current_or_former_residence ?place.
16        ?place :P87_is_identified_by ?parish .
17        ?parish :P3_has_note
       "Name-Residence"^^xsd:string .
18
19  } ORDER BY ?name
```

The code block between lines 11 and 17 of Listing 2 is designed to search for all respondents *(E21_Person)* of a given sex *(foaf:gender)* who live in a given location *(:P74_has_current_or_former_residence)*. For example it can be instantiated to, list all the female respondents living in Afurada. The property *foaf:name* describes the full name of each respondent.

For more information on the results and executed queries, please see: http://npmp.epl.di.uminho.pt.

- *Query Processor* generates the SPARQL queries according to the exhibition room requirements, sends them to the SPARQL Endpoint and after receiving the answer, combines the returned data to set up the Virtual Learning Spaces (VLS).

  We created a Python script that generates SPARQL queries according to the requirements of each exhibition room, sends them to the Fuseki (SPARQL Endpoint) and after receiving the answer, combines the data returned to configure the Virtual Learning Spaces (VLS). This Python script also includes HTML (Hyper Text Markup Language) and CSS (Cascading Style Sheets) to create and format the web page.

  Finally, to exhibit the life stories that are the objects of the Museum of the Person, the web pages were built (Virtual Learning Spaces).

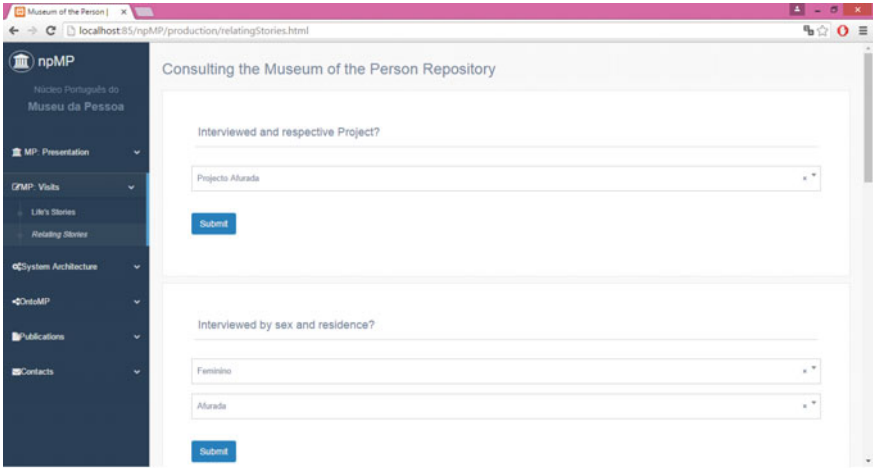
  Figure 6 displays the page where the museum visitor can perform the SPARQL queries.

  The answer to the query referred in Listing 2 (Interviewed by sex and residence) is shown in the Fig. 7.

  In this approach, each Virtual Learning Space (a museum's exhibition room) is built fulfilling a web page template with the concrete data retrieved from the data store.

4 XML Repository and Ontology Extraction

As mentioned in Sect. 3.1, we initially performed the extraction and analysis of the semantic concepts, and convert manually those triples into RDF triples. After understanding the structure of the documents, the information they contain and how to convert the ontology triples into RDF triples, we decided to develop a text filter able



Consulting the Museum of the Person Repository

Interviewed and respective Project?

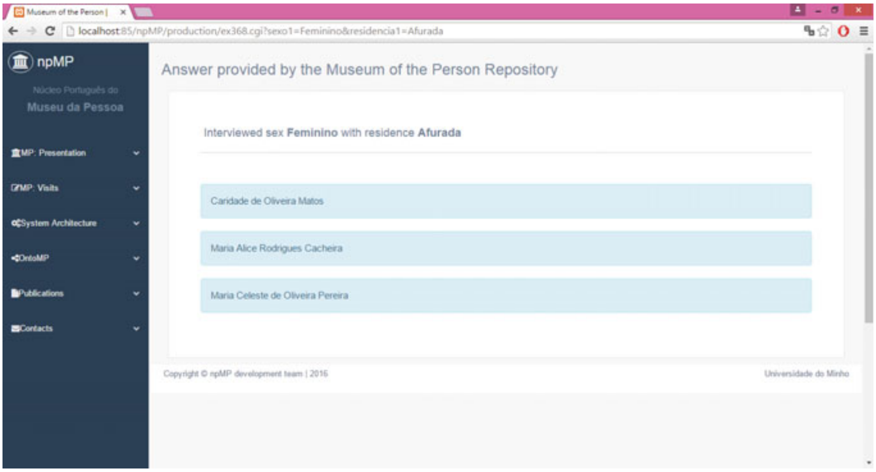
Projecto Afurada

Interviewed by sex and residence?

Feminino

Afurada

Fig. 6 Consulting the Museum of the person repository



Answer provided by the Museum of the Person Repository

Interviewed sex **Feminino** with residence **Afurada**

Caridade de Oliveira Matos

Maria Alice Rodrigues Cacheira

Maria Celeste de Oliveira Pereira

Copyright © npMP development team | 2016

Universidade do Minho

Fig. 7 Response to the SPARQL query: *Interviewed by sex and residence*

to scan all files that compose an interview (BI, Edited Interview and Photograph Captions), extract relevant information and convert into a single RDF triples file.

To develop the text filter we used the compilers generator system AnTLR (Another Tool for Language Recognition) [16] integrated in AnTLRWorks tool, version 2.1, a plugin for NetBeans IDE. AnTLR is a powerful parser generator for reading, and processing structured text files; so it is extensively used to build language-based tools, and frameworks.[3]

[3] http://www.antlr.org/index.html.

In our case, from a set of Regular Expressions (RE), AnTLR will generate a Lexical Analyzer that realizes the desired text filter for data extraction.

That text filter, or extractor, will accept an input document, like the one exemplified in Listing 3, and after analyzing and processing it will output a RDF description, like the one shown in Listing 4.

Listing 3 An XML input document

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<fotos>
    <foto ficheiro="090-F-01.jpg">
        <quem>Maria Alice Rodrigues Cacheira</quem>
        <onde>Junta de Freguesia da Afurada</onde>
        <quando data="2001-07-12"/>
    </foto>
    <foto ficheiro="090-F-02.jpg">
        <quem>Maria Alice Rodrigues Cacheira</quem>
        <onde>Junta de Freguesia da Afurada</onde><
        <quando data="2001-07-12"/>
    </foto>
</fotos>
```

Listing 4 An RDF output document

```
<rdf:Description rdf:about="&ecrm;090-F-01.jpg">
    <rdf:type rdf:resource="&ecrm;E41_Appellation"/>
</rdf:Description>

<rdf:Description rdf:about="&ecrm;I0_Interviewed_1"/>
    <rdf:type rdf:resource="&ecrm;E38_Image"/>
    <rdf:type rdf:resource="&foaf;Image"/>
    <foaf:depicts rdf:resource="&ecrm;Interviewed_1"/>
    <P67_refers_to rdf:resource="&ecrm;I0_Description_Interview_1"/>
    <P1_is_identified_by rdf:resource="&ecrm;090-F-01.jpg"/>
    <P4_has_time-span rdf:resource="&ecrm;TS1"/>
    <P7_took_place_at rdf:resource="&ecrm;PL1"/>
</rdf:Description>

<rdf:Description rdf:about="&ecrm;2001-07-12">
    <rdf:type rdf:resource="&ecrm;E49_Time_Appellation"/>
</rdf:Description>

<rdf:Description rdf:about="&ecrm;TS1">
    <rdf:type rdf:resource="&ecrm;E52_Time-Span"/>
    <P78_is_identified_by rdf:resource="&ecrm;2001-07-12"/>
</rdf:Description>

<rdf:Description rdf:about="&ecrm;PL1">
    <rdf:type rdf:resource="&ecrm;E53_Place"/>
    <P87_is_identified_by rdf:resource="&ecrm;Place1"/>
</rdf:Description>

<rdf:Description rdf:about="&ecrm;Place1">
    <rdf:type rdf:resource="&ecrm;E48_Place_Name"/>
    <P3_has_note rdf:datatype="&xsd;string">Junta de Freguesia da Afurada</P3_has_note>
</rdf:Description>

<rdf:Description rdf:about="&ecrm;I0_Description_Interview_1">
    <rdf:type rdf:resource="&ecrm;E55_Type"/>
    <P2_has_type rdf:resource="&ecrm;Description"/>
    <P3_has_note rdf:datatype="&xsd;string">Maria Alice Rodrigues Cacheira</P3_has_note>
</rdf:Description>

<rdf:Description rdf:about="&ecrm;090-F-02.jpg">
    <rdf:type rdf:resource="&ecrm;E41_Appellation"/>
</rdf:Description>

<rdf:Description rdf:about="&ecrm;I1_Interviewed_1"/>
    <rdf:type rdf:resource="&ecrm;E38_Image"/>
    <rdf:type rdf:resource="&foaf;Image"/>
    <foaf:depicts rdf:resource="&ecrm;Interviewed_1"/>
    <P67_refers_to rdf:resource="&ecrm;I1_Description_Interview_1"/>
    <P1_is_identified_by rdf:resource="&ecrm;090-F-02.jpg"/>
    <P4_has_time-span rdf:resource="&ecrm;TS1"/>
    <P7_took_place_at rdf:resource="&ecrm;PL1"/>
</rdf:Description>
```

That automatic transformation is obtained using a specification (an *AnTLR Lexer grammar*) illustrated in Listing 5. The fragment shown in the referred listing is a sequence of transformation rules that corresponds to the beginning of the global specification (the specification part not included will be discussed below). Each rule has a name and a pair composed of a Regular Expression (RE) and a Semantic Action (SE) written in Java. A rule is interpreted from left to right: if the Regular Expression is found in the input, then the corresponding Semantic Action is triggered. The RE defines the text pattern that shall be find in the input, and the SE specifies how the concrete text found shall be transformed.

Moreover, AnTLR lets the programmer to set up modes that group the specific rules to address each sub block in the input file.

In Listing 5 it can be seen the three rules (namely, *Cabec, Fotos e MP*) corresponding to the three input files (*BI, Photography Captions,* and *Edited Interview*), respectively. When the extractor reads a XML tag definning the beginning of one of these three documents, it enters a special AnTLR mode to process that document's content.

Listing 6 shows the main mode to process the *Photography Caption* XML documents. The listing illustrates the general approach adopted: when a block opening tag is found, the appropriate mode is entered to consume the block contents; when the block closing tag is found, the processor exits the mode and returns to the initial mode.

The four auxilairy modes, called from the main one (see lines 13–19), contains the specific rules used to extract information from the four main blocks of the *Photography Captions* input document. Listing 7 contains the rules (just a fragment is shown) executed at the end of the processing (mode activated at line 21) to print out the RDF triples built in the internal representation. This grammar fragment is actually responsible for the generation of the RDF output file.

Listing 5 XML2RDF Lexer Grammar for AnTLR

```
lexer grammar XML2RDF;

Cabec   :  '<'[Bb][Ii]'>'                   -> mode(sBI)
        ;
Fotos   : '<'[Ff][Oo][Tt][Oo][Ss]'>'        -> mode (sFOTOS)
        ;
MP      : '<'[Mm][Pp]'>'                    -> mode (sMP)
        ;
Default:    .          { ; }
        ;

        ...
........Modes specification........
        ...
```

Listing 6 Lexer Grammar Photos main Mode

```
mode sFOTOS;
GetSFOTOS         :  '<foto'      -> mode(sFOTO)
                    ;
OutFOTOSSAVE      :  '</fotos>'   -> mode(DEFAULT_MODE)
                 ;
DefaultsFOTOS    :    .    { ; }
                 ;


mode sFOTO;
GetFOTO      :    [ ]+'ficheiro="'     -> mode (sFICHEIRO)
             ;
GetQUEM      :   '<quem>'             -> mode (sQUEM)
             ;
GetQUANDO    :   '<quando'            -> mode (sQUANDO)
             ;
GetFACTO     :   '<facto>'            -> mode (sFACTO)
             ;
GetONDE      :   '<onde>'             -> mode (sONDE)
             ;
OutFOTOS     :   '</'                 -> mode (sPRINTTUDO)
             ;

DefaultsFOTO   :    .    { ; }
               ;
```

Listing 7 Lexer Grammar Print Mode

```
mode sPRINTTUDO;
GetsPRINTTUDO   :'foto'    {

     pessoa.AddImage("I"+newCountKeyFicheiro+"_Interviewed_"+
     countinterview);

                                 System.out.print("<rdf:Description
      rdf:about=\"&ecrm;");

     System.out.println(ficheiro+"\">");
                                 System.out.println("\t<rdf:type
     rdf:resource=\"&ecrm;E41_Appellation\"/>");

     System.out.println("</rdf:Description>\n\n");

     System.out.println("<rdf:Description
     rdf:about=\"&ecrm;I"+newCountKeyFicheiro+"_Interviewed_"+
     countinterview+"\"/>");
                                 System.out.println("\t<rdf:type
     rdf:resource=\"&ecrm;E38_Image\"/>");
                                 System.out.println("\t<rdf:type
     rdf:resource=\"&foaf;Image\"/>");

     System.out.println("\t<foaf:depicts
     rdf:resource=\"&ecrm;Interviewed_"+countinterview+"/>");

     if(!quem.equals("")){
              System.out.println("\t<P67_refers_to
     rdf:resource=\"&ecrm;I"+newCountKeyFicheiro+"_Description_I

nterview_"+ countinterview+"\"/>");}

     if(!facto.equals("")){
              System.out.println("\t<P3_has_note
     rdf:datatype=\"&xsd;string\">"+facto+"</P3_has_note>"); }

     System.out.println("\t<P1_is_identified_by
     rdf:resource=\"&ecrm;"+ficheiro+"\"/>");


         ...


OutsPRINTTUDO  :  '>'     -> mode(sFOTOS)
                  ;
```

5 Conclusion

This paper describes the creation of a virtual museum to exhibit people's life stories, called the Museum of the Person (MP). Museum of the Person[4] was born in Brazil, São Paulo, in 1991, created by a group of historians who decided to build the country's history using testimonials of ordinary people [17]. Our work concerns the

[4] Accessible at: http://www.museudapessoa.net.

Portuguese branch of such network of life stories museums, npMP. From the life stories of individuals, the objective is to write up the stories of families, communities, or institutions.

After analyzing the documents that make up the repository, we designed OntoMP, an ontology for the Museum of the Person. The next stage after the validation and tuning of OntoMP was to describe it in a standard ontology format used for museums, CIDOC-CRM (CIDOC Conceptual Reference Model) complemented with some pertinent FOAF and DBpedia concepts and properties.

In this paper we propose a general architecture to build a software platform to create the museum's virtual exhibition rooms, as web pages, extracting information from the museum's repository. To implement the overall architecture outlined there are two possible alternative techniques. However, to refine this architecture, we chose approach 1. One approach uses a TripleStore to archive the ontology instances and resorts to SPARQL technology to query the repository and obtain the information that will be exhibited. The other approach uses a Relational Database as archive and reuses CaVa framework to extract and display the information. CaVa is a novel proposal under development in the context of the PhD project of one of the authors, and our first objective was to use npMP as a second case study to test that framework.

After implementing the approach 1, we came to the conclusion that to implement the first module (Ingestion Function) manually is a very lengthy process. So we decided to create a text filter to perform the three phases of this module automatically, as was discussed in the article.

As future work we intend to refine the filter in some aspects, particularly in the recursive episodes, among others in order to be possible to deal with all the documents stored in our present repository.

Acknowledgements This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope:UID/CEC/00319/2013. The work of Ricardo Martini is supported by CNPq, grant 201772/2014-0.

References

1. Almeida, J.J., Rocha, J.G., Henriques, P.R., Moreira, S., Simões, A.: Museu da Pessoa–arquitectura. In: Encontro Nacional da Associação de Bibliotecários, Arquivista e Documentalistas, ABAD'01. BAD (2001)
2. Rodrigues, B.C., Crippa, G.: Novas Propostas e Desafios Das Mediações Culturais em Museus Virtuais. In: El Pensamiento Museuloógico Contenporá neo. O Pensamento Museulógico Contemporâneo, pp. 599–608. ICOM (2011)
3. Philip, B.: Stafford. Museum of person, Technical report (2015)
4. Araújo, C.: An Ontology for the Museum of the Person Combining CIDOC-CRM with FOAF. Universidade do Minho, Msc pre-thesis (2016)
5. Martini, R.: Formal Description and Automatic Generation of Learning Spaces based on Ontologies. Universidade do Minho, Ph.D. pre-thesis (2015)
6. Simões, A., Almeida, J.J.: Histórias de Vida + Processamento Estrutural = Museu da Pessoa. In: XATA 2003 — XML: Aplicações e Tecnologias Associadas, pp. 16. Braga, Portugal (2003). UM

7. Martini, R.G., Araújo, C., Almeida, J.J., Henriques, P.R.: New advances in information systems and technologies: volume 2. In: chapter OntoMP, An Ontology to Build the Museum of the Person, pp. 653–661. Springer International Publishing, Cham (2016)
8. Araújo, C., Martini, R.G., Henriques, P.R., Almeida, J.J.: Architectural approaches to build the museum of the person. In: Rocha, Á., Reis, L.P., Cota, M.P., Suárez, O.S., Gonçalves, R. (eds.) Sistemas y Tecnologías de Información—Atas da 11^{a} Conferência Ibérica de Sistemas e Tecnologias de Informação, volume Vol. I — Artículos de la Conferencia, pp. 383–388. AISTI–Associação Ibérica de Sistemas e Tecnologias de Informação, June 2016
9. Martini, R.G., Araújo, C., Librelotto, G.R., Henriques, P.R.: New advances in information systems and technologies. In: chapter A Reduced CRM-Compatible Form Ontology for the Virtual Emigration Museum, pp. 401–410. Springer International Publishing, Cham (2016)
10. ICOM/CIDOC. Definition of the CIDOC Conceptual Reference Model. Technical report, ICOM/CIDOC, May 2015
11. Allemang, D., Hendler, J.: Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL. Elsevier Science (2011)
12. Al-Mukhtar, M.M.A., Al-Assafy, A.T.A.: The implementation of foaf ontology for an academic social network. Int. J. Sci. Eng. Comput. Technol. **4**(1), 10 (2014)
13. Dbpedia. Ontology. http://wiki.dbpedia.org/ (2016). Accessed 15 June 2016
14. APACHE JENA. TDB. https://jena.apache.org/documentation/tdb/index.html (2016). Accessed 01 June 2016
15. APACHE JENA. Apache Jena Fuseki. https://jena.apache.org/documentation/fuseki2/index.html (2016). Accessed 01 June 2016
16. ANTLR. ANTLR. http://www.antlr.org/ (2016). Accessed 14 Sept 2016
17. Worcman, K.: The museum of the person. In: Virtual Museums, vol. 57, no. 3. ICOM (2004)